

ИНСТИТУТ ТОЧНОЙ МЕХАНИКИ И ВЫЧИСЛИТЕЛЬНОЙ ТЕХНИКИ
АКАДЕМИИ НАУК СССР

Е. А. Волков

НОВЫЕ ФОРМУЛЫ
ДЛЯ ВЫЧИСЛЕНИЯ
ЭЛЕМЕНТАРНЫХ ФУНКЦИЙ
НА БЭСМ

МОСКВА - 1957

ИНСТИТУТ ТОЧНОЙ МЕХАНИКИ И ВЫЧИСЛИТЕЛЬНОЙ ТЕХНИКИ
АКАДЕМИИ НАУК СССР

Е.А. Волков

НОВЫЕ ФОРМУЛЫ ДЛЯ ВЫЧИСЛЕНИЯ ЭЛЕМЕНТАРНЫХ
ФУНКЦИЙ НА БЭСМ

Москва - 1957

НОВЫЕ ФОРМУЛЫ ДЛЯ ВЫЧИСЛЕНИЯ ЭЛЕМЕНТАРНЫХ
ФУНКЦИЙ НА БЭСМ

В работе даются формулы с численными значениями коэффициентов для вычисления элементарных функций на БЭСМ. Эти формулы обеспечивают высокую точность вычисления с минимальной затратой машинного времени. Приведенные формулы и методика их построения могут быть использованы при разработке стандартных подпрограмм для вычислительных машин других типов.

NEW FORMULAS FOR COMPUTING ELEMENTARY
FUNCTIONS USING THE BESM

Some formulas with numerical values of factors are given for computing elementary functions using the BESM. These formulas provide high accuracy of computation as well as minimum computer time. The formulas and methods of their constructing can be used in developing standard subroutines for other computer types.

НОВЫЕ ФОРМУЛЫ ДЛЯ ВЫЧИСЛЕНИЯ ЭЛЕМЕНТАРНЫХ
ФУНКЦИЙ НА БЭСМ¹

В диодном запоминающем устройстве БЭСМ располагаются постоянные подпрограммы для перевода чисел из десятичной системы счисления в двоичную и обратно, а также для вычисления часто встречающихся в расчетах элементарных функций: $\ln x$, e^x , $\operatorname{tg} x$, $\sin x$ и $\cos x$, $\operatorname{arctg} x$, $\operatorname{arcsin} x$ и \sqrt{x} .

Ранее существовавшие подпрограммы были составлены главным образом, исходя из условий минимума занимаемого ими места. Поэтому часть ячеек диодного запоминающего устройства (ДЗУ) оставалась свободной, что дало возможность произвести изменение постоянных подпрограмм с целью повышения быстродействия БЭСМ.

Вновь созданные подпрограммы по количеству команд несколько длиннее прежних, однако они позволяют существенно сократить время вычислений элементарных функций, вследствие чего сокращается общее количество счетного времени, расходуемого для решения различных задач.

Обычная суммарная ошибка вычисления элементарных функций по новым подпрограммам не превосходит одну-две единицы младшего разряда цифровой части, за исключением тех случаев, о которых будет сказано особо.

Повышение скорости вычисления элементарных функций по сравнению со

¹ В работе принимал участие А. В. Лебедев.

старыми подпрограммами достигается: 1) еще большим уменьшением интервала, на котором производится собственно вычисление данной функции; 2) приближением функции на выбранном интервале наилучшим многочленом; 3) вычислением многочлена непосредственно (без циклов); 4) более рациональным использованием команд БЭСМ, существующих в настоящее время.

Рассмотрим формулы, которые были использованы в новых постоянных программах для вычисления элементарных функций.

Вычисление функции $\ln x$. - Любой число x в БЭСМ обычно представляется в виде

$$x = 2^p \cdot x_1, \quad (1)$$

где

$$\frac{1}{2} \leq x_1 < 1.$$

Старая подпрограмма вычисляла $\ln x$ по формуле

$$\ln x = p \ln 2 + \ln x_1, \quad (2)$$

где

$$\ln x = \sum_{k=0}^{\infty} \left(\frac{x_1 - 1}{x_1 + 1} \right)^{2k+1} \left(\frac{2}{2k+1} \right). \quad (3)$$

При этом вычислялась по циклу частичная сумма ряда (3), состоящая из членов, по абсолютной величине превосходящих 2^{-32} .

В новой программе вычисления $\ln x$ введены следующие изменения.

Для всех x_1 имеет место тождество

$$\ln x_1 = \ln \lambda x_1 - \ln \lambda. \quad (4)$$

При $\lambda = \sqrt{2}$ значительно ускоряется сходимость ряда (3) в самых неблагоприятных случаях. С помощью наилучшего многочлена от u вычисляется $\ln x$ с погрешностью, не превосходящей 2^{-32} , по формуле:

$$\ln x = (p - \frac{1}{2}) \ln 2 + u (b_0 + u^2 (b_1 + u^2 (b_2 + u^2 b_3))), \quad (5)$$

где

$$u = \frac{x_1 - \frac{\sqrt{2}}{2}}{x_1 + \frac{\sqrt{2}}{2}} \quad \text{и} \quad b_0, b_1, b_2, b_3 - \text{постоянные},$$

имеющие следующие численные значения:

$$b_0 = 1,999\ 999\ 993\ 788$$

$$b_1 = 0,666\ 669\ 470\ 507$$

$$b_2 = 0,399\ 659\ 100\ 019$$

$$b_3 = 0,300\ 974\ 506\ 336.$$

Вычисление функции e^x . - Вычисление функции e^x основывается на следующем свойстве показательной функции:

$$e^x = 2^{\frac{x}{\ln 2}} = 2^{\left[\frac{x}{\ln 2}\right]} \cdot 2^{\left\{\frac{x}{\ln 2}\right\}} = 2^{\left[\frac{x}{\ln 2}\right]} \cdot e^z, \quad (6)$$

где

$$e^z = e^{\ln 2 \left\{\frac{x}{\ln 2}\right\}}. \quad (7)$$

В старой программе для вычисления e^z учитывались 12 членов ряда

$$e^z = \sum_{k=0}^{\infty} \frac{z^k}{k!}, \quad (8)$$

вычисление которых осуществлялось при помощи цикла.

При составлении новой программы вычисления функции e^z было использовано соотношение

$$e^z = \sqrt{2} \cdot e^{z - \frac{\ln 2}{2}} = \sqrt{2} \cdot e^u, \quad (9)$$

где

$$-\frac{\ln 2}{2} \leq u = z - \frac{\ln 2}{2} < \frac{\ln 2}{2}. \quad (10)$$

Функция e^u приближается наилучшим многочленом седьмой степени. Окончательное выражение e^z берется от аргумента u , являющегося дробной частью величины $\frac{x}{\ln 2}$

$$0 \leq u = \left\{ \frac{x}{\ln 2} \right\} < 1.$$

В пределах требуемой точности (2^{-33}) e^z имеет выражение

$$e^z = c_0 + c_1 u + c_2 u^2 + c_3 u^3 + c_4 u^4 + c_5 u^5 + c_6 u^6 + c_7 u^7. \quad (11)$$

Значения коэффициентов этого многочлена следующие:

$$\begin{aligned}c_0 &= 0, 999\ 999\ 999\ 93 \\c_1 &= 0, 693\ 147\ 187\ 87 \\c_2 &= 0, 240\ 226\ 356\ 70 \\c_3 &= 0, 055\ 505\ 295\ 42 \\c_4 &= 0, 009\ 613\ 530\ 02 \\c_5 &= 0, 001\ 342\ 985\ 66 \\c_6 &= 0, 000\ 142\ 992\ 74 \\c_7 &= 0, 000\ 021\ 651\ 59.\end{aligned}$$

Ввиду того, что в БЭСМ операция умножения требует приблизительно в три с половиной раза больше времени, чем сложение или вычитание, для вычисления полученного многочлена используется следующая формула, сокращающая число необходимых умножений:

$$e^x = c_0 + \bar{x} \{ a_1 + \bar{x} [a_2 + \bar{x} \{ [(\bar{x} + B)^2 + C + \bar{x}] [(\bar{x} + B)^2 + D] - E \}] \}, \quad (12)$$

где

$$\begin{aligned}\bar{x} &= \lambda v; & \lambda &= \sqrt[7]{c_7}; & a_1 &= \frac{c_1}{\sqrt[7]{c_7}}; \\a_2 &= \frac{c_2}{(\sqrt[7]{c_7})^2}; & a_3 &= \frac{c_3}{(\sqrt[7]{c_7})^3}; & a_4 &= \frac{c_4}{(\sqrt[7]{c_7})^4}; \\a_5 &= \frac{c_5}{(\sqrt[7]{c_7})^5}; & a_6 &= \frac{c_6}{(\sqrt[7]{c_7})^6}; \\B &= \frac{a_6 - 1}{4}; & C &= (1 + 2B)(a_5 - 2B - 6B^2) - (a_4 - B^2 - 4B^3); \\D &= (a_4 - B^2 - 4B^3) - 2B(a_5 - 2B - 6B^2); \\E &= B^4 + B^2(C + D) + CD - a_3.\end{aligned}$$

Коэффициенты этого многочлена имеют следующие значения:

$$\lambda = 0, 215\ 596\ 346\ 446$$

$$a_1 = 3, 215\ 022\ 885\ 576$$

$$a_2 = 5, 168 182 735 768$$

$$B = 0, 105 963 619 947$$

$$C = -1, 277 917 410 482$$

$$D = 3, 881 751 544 667$$

$$E = -10, 469 925 626 182 .$$

Просчет контрольных значений e^z с использованием этих коэффициентов дает совпадение одиннадцати значащих цифр.

При вычислении функции e^x , кроме обычной ошибки округления, добавляется потеря числа двоичных разрядов цифровой части, приблизительно равная числу двоичных разрядов, которыми представляется модуль целой части $\lfloor \frac{x}{\ln 2} \rfloor$ данной функции e^x .

Вычисление функций $\sin x$ и $\cos x$. - Для вычисления функций $\sin x$ и $\cos x$ по старым программам использовалось разложение $\tg x$ в непрерывную дробь, которая вычислялась циклом, состоящим из пяти команд.

Вычисление $\sin x$ и $\cos x$ по новой программе осуществляется с помощью приведения аргумента x к отрезку $0 \leq x_1 \leq \frac{\pi}{4}$.

Функции $\sin x$ и $\cos x$ находятся через тангенс половины приведенного аргумента.

Для вычисления $\tg \frac{x_1}{2}$ используется разложение $\ctg \frac{w}{2}$ в ряд Лорана

$$\ctg \frac{w}{2} = \frac{1}{w} - \sum_{k=1}^{\infty} \frac{2^{2k} |B_{2k}|}{(2k)!} w^{2k-1}, \quad (13)$$

откуда

$$2\tg \frac{x_1}{2} = \frac{x_1}{1 - \sum_{k=1}^{\infty} \frac{|B_{2k}|}{(2k)!} x_1^{2k}}. \quad (14)$$

Для аппроксимации знаменателя используется наилучший многочлен четвертой степени от z^2 .

$$2\tg \frac{x_1}{2} = \frac{z}{c_0 - c_1 z^2 - c_2 z^4 - c_3 z^6 - c_4 z^8}, \quad (15)$$

где $z = \frac{x_1}{\pi} - \frac{4}{\pi}$ — значение дробной части, получаемое на БЭСМ при приведении аргумента к отрезку $[0, \frac{\pi}{4}]$.

Вычисляя, получаем следующие значения коэффициентов:

$$c_0 = 1, 273 239 544 731$$

$$c_1 = 0, 065 449 846 718$$

$$c_2 = 0, 000 672 881 123$$

$$c_3 = 0, 000 009 877 325$$

$$c_4 = 0, 000 000 158 336 .$$

В целях сокращения времени вычисления многочлена, стоящего в знаменателе формулы (15), применяется не схема Горнера, требующая четыре операции сложения и пять умножений, а более экономная, в которой используются 4 умножения и 5 сложений.

Используя подстановку $\bar{x} = \lambda z$, получаем из формулы (15) окончательную формулу для вычисления:

$$2 \operatorname{tg} \frac{x_1}{2} = \frac{\bar{x}}{E - [(\bar{x}^2 + B)^2 + C + \bar{x}^2] [(\bar{x}^2 + B)^2 + D]} , \quad (16)$$

где коэффициенты B , C , D и E определяются по формулам:

$$B = \frac{a_3 - 1}{4} ,$$

$$C = (1 + 2B)(a_2 - 2B - 6B^2) - (a_1 - B^2 - 4B^3) ,$$

$$D = (a_1 - B^2 - 4B^3) - 2B(a^2 - 2B - 6B^2) ,$$

$$E = B^4 + B^2 (C + D) + CD + a_0 ,$$

причем

$$a_0 = \lambda c_0 ; \quad a_1 = \frac{c_1}{\lambda} ; \quad a_2 = \frac{c_2}{\lambda^2} ; \quad a_3 = \frac{c_3}{\lambda^3} \quad \text{и} \quad \lambda = \sqrt[7]{c_4} .$$

Значения коэффициентов, входящих в формулу (16), будут следующие:

$$\lambda = 0,106 785 251 669$$

$$B = -0,072 162 649 192$$

$$C = -0,039\ 607\ 473\ 057$$

$$D = 0,705\ 279\ 988\ 224$$

$$E = 0,111\ 522\ 419\ 569$$

Просчет контрольных значений $\sin x$ и $\cos x$ с использованием приведенных коэффициентов дает совпадение одиннадцати десятичных знаков.

Следует заметить, что при приведении аргумента к отрезку $0 \leq x_1 \leq \frac{\pi}{4}$ происходит потеря точности значения функции, равная n - двоичным разрядам, где n - двоичный порядок числа, полученного при делении исходного аргумента x на $\frac{\pi}{4}$. Эта ошибка в полученном значении функции является принципиальной и ее избежать практически нельзя.

Вычисление квадратного корня. - По старой программе вычисление значений функции $y = \sqrt{x}$ велось циклом по итерационной формуле

$$y_{n+1} = \frac{y_n + \frac{x}{y_n}}{2} \quad (17)$$

При этом за начальное приближение бралась величина

$$y_0 = 2^{\left[\frac{n}{2}\right]} \text{sign } n, \quad (18)$$

где n - двоичный порядок числа x .

Вычисление квадратного корня по новой программе ведется следующим образом.

Рассматриваются два случая:

$$1) \quad x = 2^{2p} x_1, \quad (19)$$

тогда

$$y = 2^p \sqrt{x_1}. \quad (20)$$

Для вычисления $\sqrt{x_1}$ находится кусочно-линейным способом начальное приближение y_0 , как показано на рис.1.

Значение величин κ , b и Δ взяты следующие:

$$\kappa = 0,57155$$

$$b = 0,75787$$

$$\Delta = 0,013857$$

Тогда для вычисления $\sqrt{x_1}$ с точностью до 2^{-33} требуются только две итерации по формуле (17), которые сворачиваются в единую формулу

$$y = \frac{1}{4} \left(y_0 + \frac{x_1}{y_0} \right) + \frac{x_1}{\left(y_0 + \frac{x_1}{y_0} \right)} . \quad (21)$$

$$2) \quad x = 2^{p+1} x_1 , \quad (22)$$

тогда

$$y = 2^p \sqrt{2x_1} . \quad (23)$$

Вычисление значения $\sqrt{2x_1}$ ведется аналогично вычислению $\sqrt{x_1}$, только в этом случае при вычислении начального приближения y_0 вместо множителя k берется множитель $\sqrt{2}k$ и в формуле (21) вместо x_1 берется $2x_1$.

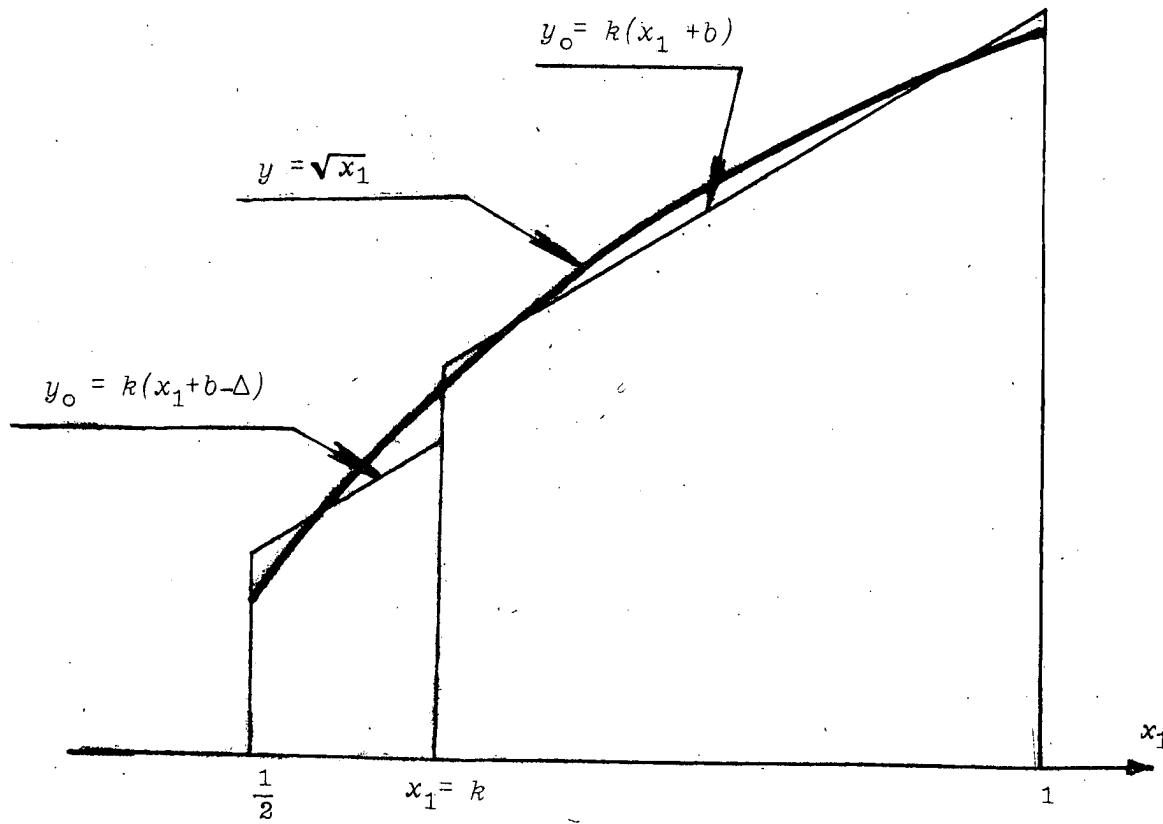


Рис. 1

Вычисление функции arctg x. - Функция $\text{arctg } x$ по старой программе вычислялась по рекуррентной формуле:

$$\text{arctg } x = \frac{\pi}{4} (\text{sign } y_1 + \frac{1}{2} \text{sign } y_2 + \dots + \frac{1}{2^k} \text{sign } y_{k+1} + \dots) , \quad (24)$$

где

$$y_1 = x \quad \text{и} \quad y_{k+1} = \frac{y_k}{2} - \frac{1}{2y_k}.$$

Приведенный выше ряд вычислялся циклом. На вычисление каждого двоичного разряда требовалось повторение цикла, состоящего из восьми команд.

В новой программе использован другой метод вычисления функции $\arctg x$. Нетрудно свести вычисление $\arctg x$ к отрезку $0 \leq x_1 \leq 1$.

Далее целесообразно вести вычисления по формуле:

$$\arctg x_1 = \arctg z + C \left\{ \begin{array}{l} x_1 < x_0, \quad z = x_1, \quad c = 0 \\ x_1 \geq x_0, \quad z = \frac{x_1 - \frac{1}{\sqrt{3}}}{1 + \frac{x_1}{\sqrt{3}}}; \quad c = \arctg \frac{1}{\sqrt{3}} = \frac{\pi}{6}, \end{array} \right. \quad (25)$$

где

$$x_0 = \frac{\sqrt{3} - 1}{\sqrt{3} + 1}.$$

Очевидно, $|z| \leq x_0 = \frac{\sqrt{3} - 1}{\sqrt{3} + 1} \approx 0,268$ (рис.2).

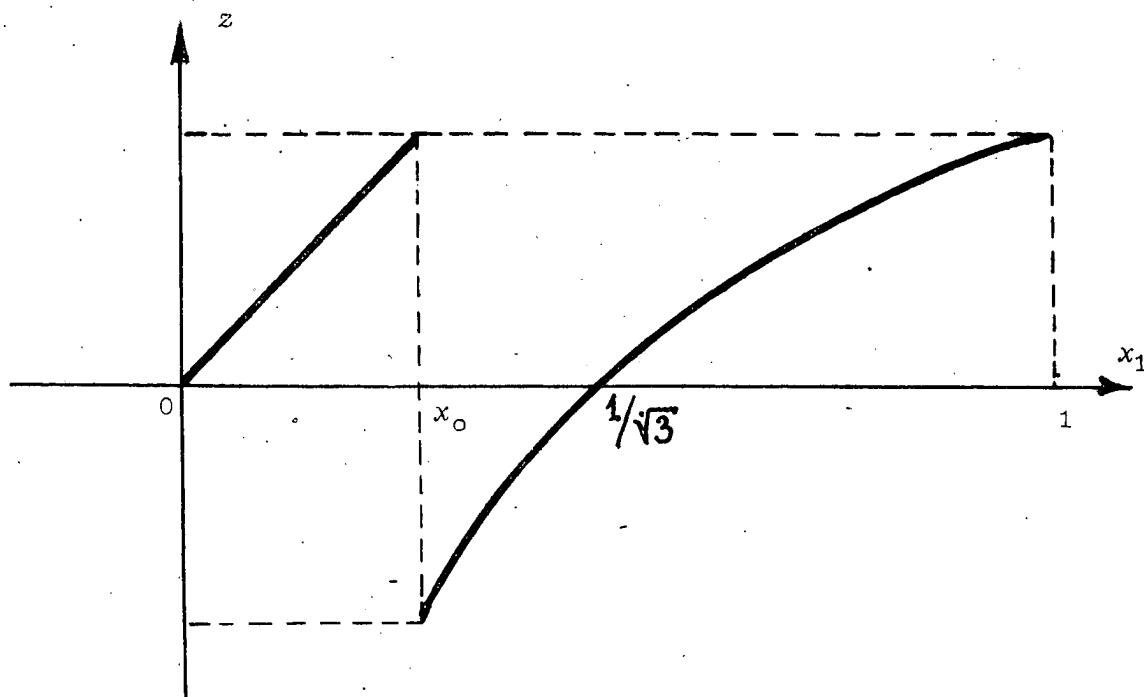


Рис. 2

В этом случае достаточно быстро сходится ряд:

$$\operatorname{arctg} z = z - \frac{z^3}{3!} + \frac{z^5}{5!} - \dots$$

Для вычисления $\operatorname{arctg} z$ в пределах требуемой точности используется следующий наилучший многочлен девятой степени:

$$\operatorname{arctg} z = l_0 z - l_1 z^3 + l_2 z^5 - l_3 z^7 + l_4 z^9, \quad (26)$$

ошибка которого не превосходит величины $\epsilon = 0,45 \cdot 10^{-10}$.

Полученный многочлен вычисляется без циклов по схеме Горнера:

$$\operatorname{arctg} z = z(l_0 - z^2(l_1 - z^2(l_2 - z^2(l_3 - l_4 z^2)))), \quad (27)$$

Коэффициенты этого многочлена имеют следующие значения:

$$\begin{aligned} l_0 &= 0,999\ 999\ 998\ 43 \\ l_1 &= 0,333\ 332\ 893\ 64 \\ l_2 &= 0,199\ 965\ 347\ 80 \\ l_3 &= 0,141\ 734\ 606\ 13 \\ l_4 &= 0,094\ 919\ 549\ 52 \end{aligned}$$

Для увеличения скорости вычисления многочлена, аппроксимирующего $\operatorname{arctg} z$, можно воспользоваться схемой более экономной, чем схема Горнера и вычислять

$$\operatorname{arctg} z = z \left\{ [(Az^2 + B)^2 + C + z^2][(Az^2 + B)^2 + D] - E \right\}, \quad (28)$$

где коэффициенты A, B, C, D и E определяются по формулам:

$$A = \sqrt[4]{l_4},$$

$$B = -\frac{l_3 + A^2}{4A^3},$$

$$C = \frac{1 + 2AB}{A^2} (l_2 - 2AB - 6A^2B^2) + (l_2 + B^2 + 4AB^3),$$

$$D = -(l_1 + B^2 + 4AB^3) - \frac{2B}{A}(l_2 - 2AB - 6A^2B^2),$$

$$E = B^4 + B^2(C + D) + CD - l_0.$$

Вычисляя, получаем следующие значения коэффициентов:

$$A = 0,555\ 058\ 703\ 74$$

$$B = -0,657\ 607\ 298\ 52$$

$$C = 0,248\ 824\ 379\ 98$$

$$D = 0,175\ 045\ 006\ 22$$

$$E = -0,586\ 132\ 618\ 27$$

При просчете контрольных значений функции $\operatorname{arctg} z$ получается совпадение десяти десятичных знаков.

Для вычисления $\operatorname{arctg} z$ такой способ вычисления многочлена на БЭСМ не был применен вследствие того, что это потребовало бы еще одну лишнюю рабочую ячейку.

Вычисление функции $\arcsin x$. — Для вычисления функции $\arcsin x$ по старой программе использовалась рекуррентная формула

$$\arcsin x = \frac{\pi}{4} (\operatorname{sign} y_1 + \frac{1}{2} \operatorname{sign} y_1 y_2 + \dots + \frac{1}{2k} \operatorname{sign} y_1 y_2 \dots y_{k+1} + \dots), \quad (29)$$

где

$$y_1 = x \quad \text{и} \quad y_{k+1} = 2y_k^2 - 1.$$

Вычисление функции $\arcsin x$ по новой программе ведется с помощью подпрограмм, вычисляющих $\operatorname{arctg} z$ и \sqrt{x} :

$$\arcsin x = \begin{cases} \operatorname{arctg} \frac{x}{\sqrt{1-x^2}}, & |x| < \frac{\sqrt{2}}{2} \\ \operatorname{arcctg} \frac{\sqrt{1-x^2}}{x}, & |x| > \frac{\sqrt{2}}{2} \end{cases} \quad (30)$$

Кроме обычной ошибки округления при вычислении функции $\arcsin x$, может быть еще потеря точности значения функции, равная потере точности при вычитании x^2 из 1. Эту потерю точности следует считать естественной, так как при x , стремящемся к единице, производная $\arcsin x$ неограниченно возрастает. Старая же программа давала потерю точности при малых значениях аргумента, что являлось ее существенным недостатком.

Таковы основные формульные изменения, которые были использованы при составлении новых постоянных подпрограмм, установленных на БЭСМ.

Новые постоянные подпрограммы для вычисления элементарных функций были введены в действие в конце сентября 1956 года. При установке постоянных программ проводились эксперименты по проверке скорости вычисления каждой из функций по старой и новой программам. Результаты этой проверки приведены ниже в таблице.

Функция	Время (в сек.) вычисления 10000 значений функции		Соотношение времени
	по новой программе	по старой программе	
$\operatorname{tg} x$, $\cos x$ *	42	75	1,8
$\sin x$ и $\cos x$	39	80,5	2,1
\sqrt{x}	19,5	41	2,1
$\ln x$	25	54,5	2,2
e^x	28	95	3,4
$\arcsin x$	53	228,5	4,3
$\operatorname{arctg} x$	29	251	8,7

При решении различных задач на БЭСМ значительное количество времени расходуется на вычисление элементарных функций. Опыт работы последнего времени показывает, что скорость решения на БЭСМ некоторых важных задач, требующих нескольких десятков часов машинного времени, повысилась на 15 - 25% за счет введения более быстродействующих программ для вычисления элементарных функций. Для отдельных задач время решения сократилось до 50%.

* Старая программа вычисляла только $\operatorname{tg} x$.

Л И Т Е Р А Т У Р А

1. Е.А. Волков. О повышении скорости вычисления элементарных функций на БЭСМ, Труды Всесоюзной конференции "Пути развития Советского математического машиностроения и приборостроения", Москва, 1956 г.
2. Todd J. Motivation for working in numerical analysis, Comm. Pure Appl. Math., 1955, vol. 8, pp.97-116.

Поступило 14/1X-1957 г.

Зак. 52

Тип. 350

ИТМ и ВТ АН СССР. Москва, Калужское шоссе, 71а